ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

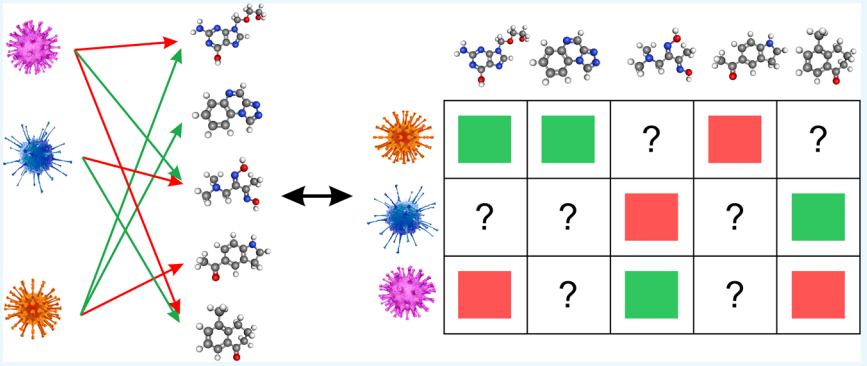
Recommender Systems in Antiviral Drug Discovery

Ekaterina A. Sosnina,* Sergey Sosnin, Anastasia A. Nikitina, Ivan Nazarov, Dmitry I. Osolodkin, and Maxim V. Fedorov

Cite This: *ACS Omega* 2020, 5, 15039−15051

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Recommender systems (RSs), which underwent rapid development and had an enormous impact on e-commerce, have the potential to become useful tools for drug discovery. In this paper, we applied RS methods for the prediction of the antiviral activity class (active/inactive) for compounds extracted from ChEMBL. Two main RS approaches were applied: collaborative filtering (Surprise implementation) and content-based filtering (sparse-group inductive matrix completion (SGIMC) method). The effectiveness of RS approaches was investigated for prediction of antiviral activity classes ("interactions") for compounds and viruses, for which some of their interactions with other viruses or compounds are known, and for prediction of interaction profiles for new compounds. Both approaches achieved relatively good prediction quality for binary classification of individual interactions and compound profiles, as quantified by cross-validation and external validation receiver operating characteristic (ROC) score >0.9. Thus, even simple recommender systems may serve as an effective tool in antiviral drug discovery.

1. INTRODUCTION

Multitask learning[1] gained popularity in different fields by virtue of continued rapid growth of available information and the development of advanced algorithms. The advantage of multitask learning is that it provides an opportunity to use additional information from related tasks for prediction for a task with insufficient information. It allows one both to improve the prediction performance and to process small or imbalanced data sets that are prevalent in the drug discovery field.[2]

The multitask learning approaches in different fields evolved separately and are characterized by different definitions and notations (see the Appendix 1). The concept of multitask learning has become popular in chemoinformatics[3−7] and the clearest examples of its realization were the proteochemometrics[8−10] and "read across" approach.[11] In e-commerce, multitask learning is referred to as the recommender system (RS).

RS is one of the approaches[12] based on multitask learning and allows one to realize multitask prediction (occasionally referred to as multitarget prediction[13]). The interest in RS began to increase in October 2006 with the announcement of a Netflix Prize competition,[14,15] aiming to create a precise system to analyze users' preferences and suggest video content for them.

RS methods are classified, based on the information used for model creation, into collaborative filtering (CF), content-based filtering (CBF), hybrid approaches, and others.[16,17] The choice of approach for a certain task is grounded on the required prediction accuracy and available computational resources.[18−20]

Collaborative filtering (CF)[16,21,22] is one of the most common RS methods, popularized during the Netflix competition due to its simplicity. If users have similar preferences, then they have similar profiles and *vice versa*. A CF RS model recommends new content to a user based on its evaluation of other users with similar preference profiles. In the drug discovery context, CF methods may rely on the similarity between compound or target interaction profiles to predict interaction values and select compound−target pairs with higher interaction scores.

Received: February 26, 2020
Accepted: June 3, 2020
Published: June 21, 2020

CF methods are the easiest to implement but have a lot of limitations.[16,23−25] The most important one is the cold-start problem (CS): interaction values cannot be reliably predicted for pairs consisting of new compounds or targets due to an inability to calculate similarity for their (empty) interaction profiles. The second limitation is the sparsity problem: the fewer the interaction values known, the more complicated it is to calculate similarity. The third limitation is scalability: the computational and memory complexities of CF algorithms are generally quadratic.

Content-based filtering (CBF) RS methods are more advanced and allow one to predict interaction values based on additional feature information, also called side-channel information, which characterizes both compounds and targets.[16,26,27] CBF recommends items similar to those liked by a user in the past based on the assessment of similarity of their features. In drug discovery, CBF may employ similarity based on features, or descriptors, of compounds or targets. Feature information allows one to overcome the disadvantages inherent to CF methods: prediction for new compounds or targets and very sparse data matrices. Also, a valuable advantage of CBF algorithms is the possibility of interpreting the model by analysis of important features. The disadvantages of CBF include the ability of overfitting and the need for feature calculation, which may be complicated in the case of target characterization.

Among the rapidly growing number of multitask prediction applications in drug discovery, only a couple of dozen studies regarded their approaches as RS. These studies were usually concerned with the analysis of approved drugs and their possible side effects,[28−30] drug repurposing,[30−33] drug−drug interactions,[34,35] toxicogenomics prediction,[36] or treatment recommendations.[37,38] A comparison of several methods demonstrated the possibility to successfully use rather simple RS algorithms in drug discovery.[39] Also, there were publications describing new methods of matrix completion with validation on drug databases,[40] and estimating their robustness on bioactivity data sets.[41] In clinical medicine, RSs have been applied since 2008 to improve treatment recommendation schemes. For example, the RS approaches were used for automatic detection of omissions in medication lists,[42,43] as well as for treatment optimization in the context of the information overload problem, by suggesting knowledge-based items of interest to clinicians for specific diseases.[44]

The search for new antivirals is an attractive field for the application of the RS approach. It is rather different from other medicinal chemistry fields because the majority of primary antiviral activity data are obtained from phenotypic antiviral assays, usually cell-based.[45] Contrary to common approaches, where targets are represented by individual proteins, antiviral activity is usually measured in much more complicated systems, containing at least viruses, cells, and compounds, and approaches based on individual targets are of limited use here. Thus, the search for broad-spectrum antivirals or antivirals against less-studied viruses reduces to the application of common molecules or privileged classes, such as nucleosides, as could be seen during the current coronavirus disease pandemic.[46−48] In our previous studies, we compiled a large annotated data set of small-molecule antiviral activity, ViralChEMBL v. 0.1.[45] After filtering, the data on activity and inactivity of approximately 250K compounds against 158 viral species were represented as a sparse matrix of compound−virus interactions, containing only 400K data points of 40M possible. Typically, the sparsity of interaction matrix $\mathbf{M}$ in the RS setting reaches 90−99%,[49,50] so this data set forms a proper base for the development of predictive models based on matrix completion methods.

In this paper, we present an attempt to apply RS approaches in the antiviral drug discovery context. To complete the antiviral activity matrix, we used the CF algorithm implemented in Surprise package[51] and sparse-group inductive matrix completion (SGIMC) implementation of CBF.[52] Several questions were addressed in our study: (1) Are the RS approaches effective in the context of antiviral activity prediction taking into account the data sparsity and unusual complexity of targets (viruses)? (2) Which RS approach gives a more accurate prediction? (3) Can we obtain a reliable prediction result for new compounds or new viral species? To address these challenges, we developed scenarios for prediction of new point interactions for compounds and viruses, which were used for model building, and prediction of interaction profiles for new compounds or viruses, not used for model building.

2. MATERIALS AND METHODS

2.1. Recommender System Approaches. *2.1.1. Collaborative Filtering.* We used CF implementation of the Surprise Python package.[51] The methods we used operate only the interaction matrix elements and can be divided into three groups:

- *k*-Nearest-neighbor (kNN)-based algorithms are implemented in *knns.KNNBasic* class and identify the neighbors for the compounds and viruses based on the similarity of their interaction profiles. We used cosine similarity and mean-squared difference metrics for similarity calculation.
- Clustering algorithms are implemented in the *co_clustering.CoClustering* class. They identify the neighborhood by grouping compounds or viruses into coclusters, simultaneously clustering the columns and rows of a matrix, and generate predictions based on the average interaction values.
- Matrix factorization algorithms are represented by singular value decomposition (*matrix_factorization.SVD*) and non-negative matrix factorization (*matrix_factorization.NMF*) methods. They are based on the idea of interaction matrix decomposition and determination of the latent variables allowing for completion of the missing interaction values.

Model hyperparameters are given in Supporting Information Table S1.

2.1.2. Content-Based Filtering. The sparse-group inductive matrix completion algorithm implemented in the SGIMC package[52] was used as an example of CBF RS. It is based on the inductive matrix completion (IMC) method and allows one to filter out noninformative features. To recover the missing entries of matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$, where only M_{ij} for $(i, j) \in \Omega \subset \{1,...,n_1\} \times \{1,...,n_2\}$ are known, IMC takes into account side feature information provided in matrices $\mathbf{X}$ and $\mathbf{Y}$. In our case, $\mathbf{X} \in \mathbb{R}^{n_1 \times d_1}$ contains descriptors of the compounds and $\mathbf{Y} \in \mathbb{R}^{n_2 \times d_2}$ contains virus features (here, taxa), n_1 and n_2 are the numbers of compounds and species, and d_1 and d_2 are the numbers of their features, respectively.

The approach is based on the assumption that the elements of the matrix $\mathbf{M}$ may be predicted via a bilinear model: $M_{ij} \sim \mathbf{x}_i^T \mathbf{W} \mathbf{y}_j$ for a low-rank matrix $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$, given that the features are predictive. As the matrix $\mathbf{W}$ must have rank $k < \min(d_1; d_2)$,

https://dx.doi.org/10.1021/acsomega.0c00857
ACS Omega 2020, 5, 15039−15051

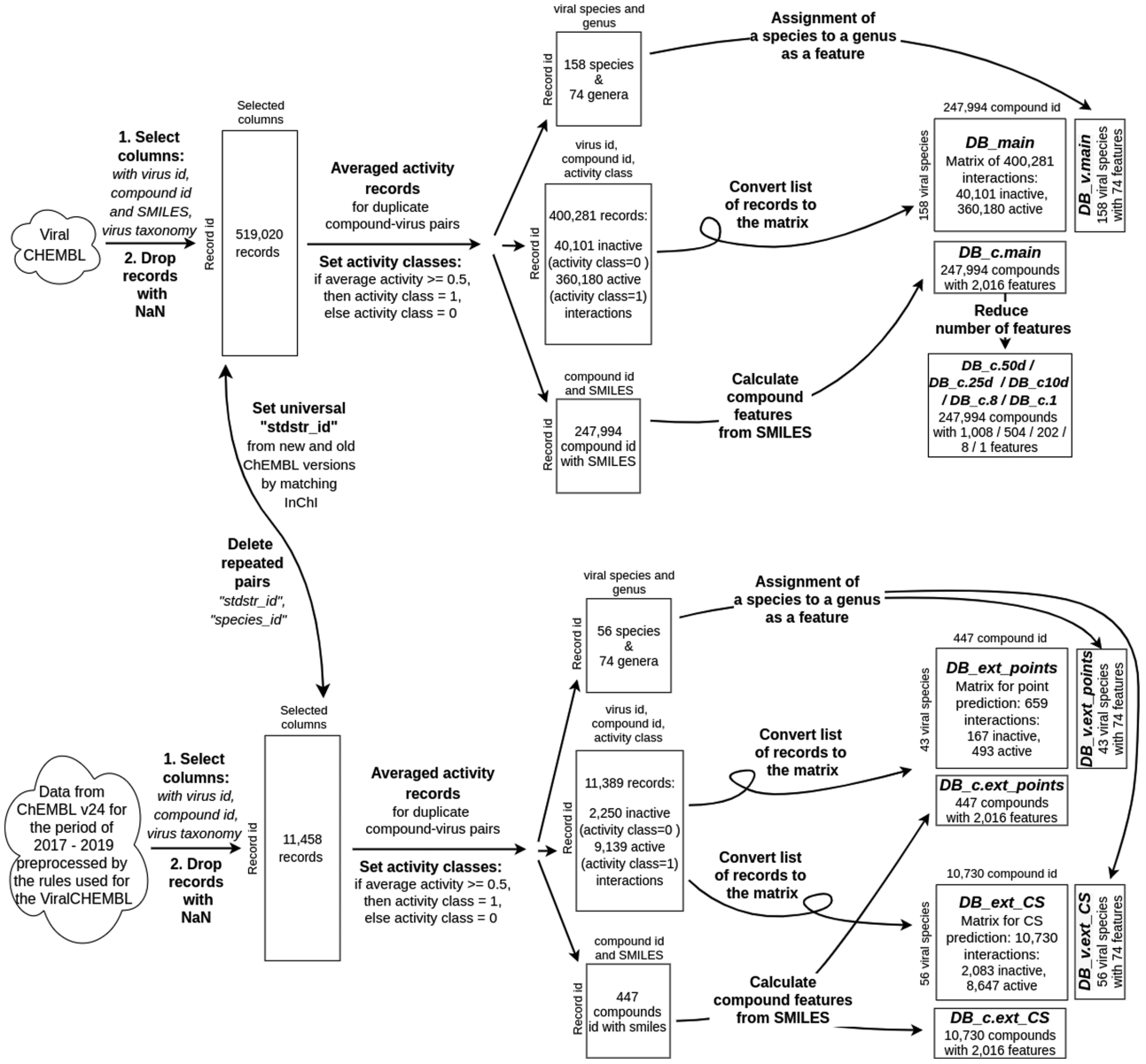


Figure 1. Scheme of data preparation.

according to the constraint of the inductive matrix completion approach, the matrix $\mathbf{W}$ can be represented by a low-rank product $\mathbf{UV}^T$ with $\mathbf{U} \in \mathbb{R}^{d_1 \times k}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times k}$. Then, the penalized minimization problem is solved

$$\min_{U,V} \sum_{(i,j)\in\Omega} \mathcal{L}(M_{ij}, (\mathbf{XUV}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}})_{ij}) + \lambda_U R(\mathbf{U}) + \lambda_V R(\mathbf{V}) \tag{1}$$

where $\mathcal{L}$: $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the smooth convex loss function for the binary classification problem, $\log(1 + e^{-yp})$ for y and p as known and predicted values, respectively. $R(\cdot)$ is the penalty, and λ_V and λ_U are the appropriate regularization coefficients.

The SGIMC algorithm shares the idea of the IMC approach of matrix completion by combining feature vectors associated with rows and columns of an interaction matrix with a low-rank matrix. The method differs by application of the sparse-group penalty for selection of side features, in addition to the classic ridge and lasso regularizations used in IMC. Thus, the penalty function $R(\cdot)$ is represented as a sum of three terms: sparsity-inducing penalty $\|\mathbf{Z}\|_{2,1}$, the squared Frobenius norm $\|\mathbf{Z}\|_{\mathrm{F}}^2$, and the matrix L_1-norm $\|\mathbf{Z}\|_{1,1}$

$$\lambda_Z R(\mathbf{Z}) = C_{\mathrm{lasso}}\|\mathbf{Z}\|_{1,1} + C_{\mathrm{ridge}}\|\mathbf{Z}\|_{\mathrm{F}}^2 + C_{\mathrm{group}}\|\mathbf{Z}\|_{2,1} =$$

$$= C_{\mathrm{lasso}} \sum_{i=1}^{n}\sum_{j=1}^{d} |z_{ij}| + C_{\mathrm{ridge}} \sum_{i=1}^{d}\sum_{j=1}^{k} z_{ij}^2 + C_{\mathrm{group}} \sum_{i=1}^{d} \|e_i^{\mathrm{T}}\mathbf{Z}\|_2 \tag{2}$$

where $\mathbf{e}_i$ is an ith unit vector, which conforms to the dimensionality of its context. The algorithm relies on single penalty functions or their combinations by setting the proper regularization coefficients C_{lasso}, C_{ridge}, and C_{group}.

We investigated the influence of regularization coefficients, the rank of low-rank matrix $\mathbf{W}$, and the number of training iterations on the predictive ability of RS in the case of antiviral activity data. The ranges of the investigated hyperparameter values are provided in Supporting Information Table S1.

2.2. Data Preparation. We used ViralChEMBL[45] as the source of information about compound−virus interactions to

https://dx.doi.org/10.1021/acsomega.0c00857
ACS Omega 2020, 5, 15039−15051

create data sets for cross-validation and model training. The ViralChEMBL data set contains 615 029 antiviral activity data points extracted from ChEMBL v.20, standardized and annotated by virus species according to ICTV taxonomy. We prepared compound−virus interaction data based on the workflow depicted in Figure 1 and described in the Supporting Information. As a result of the data processing, the *DB_main* data set for training and cross-validation comprised 247 994 compounds, 158 viral species, and 400 281 interaction values.

External validation was based on compound−virus interaction information from ChEMBL v.24 for the period 2017−2019.[53] Test sets were prepared for assessment of the interaction prediction quality in two cases: for known compounds and viruses (*DB_ext_points*) and for new compounds—compoundwise CS prediction (*DB_ext_CS*). *DB_ext_points* consisted of 659 interactions between 447 compounds and 43 viral species, while *DB_ext_CS* contained 10 730 interactions between 10 730 compounds and 55 viral species. The descriptive statistics of the data sets are given in Table 1.

Table 1. Data Sets for Model Training, Cross-Validation, and External Validation

interaction data sets	*DB_main*	*DB_ext_points*	*DB_ext_CS*
number of compounds	247 994	447	10 730
number of viral species	158	43	56
number of interactions	400 281	659	10 730
active/inactive class ratio	9/1	3/1	4/1
minimum/average/maximum number of interactions with viruses for each compound	1/1.61/36	1/1.47/12	1/1/1
minimum/average/maximum number of interactions with compounds for each virus	1/2533.42/85 823	1/15.33/155	1/195.09/2621
sparsity	1.02%	3.43%	1.82%
virus feature sets	*DB_v.main*	*DB_v.ext_points*	*DB_v.ext_CS*
number of species features	74	74	74
compound feature sets	*DB_c.main/DB_c.50d/DB_c.25d/DB_c.10d/DB_c.8/DB_c.1*	*DB_c.ext_points*	*DB_c.ext_CS*
number of compound features	2016/1008/504/202/8/1	2016	2016

2.2.1. Compound Features. Two-dimensional descriptors of chemical structures were calculated with Dragon 7.0.8[54] and used as the features of the compounds. Descriptors were calculated and selected with default options, except the following set "on": (1) exclude descriptors with constant and near-constant values, (2) exclude descriptors with a standard deviation (SD) of <0.0001, (3) exclude descriptors with all missing values, and (4) round descriptor values. The features with "NaN" values were dropped according to the SGIMC requirements. The selected features were standardized with the *StandardScaler* class of the *sklearn.preprocessing* module. The resulting compound feature matrices *DB_c.main*, *DB_c.ext_points*, and *DB_c.ext_CS* were composed of 2016 feature columns and different numbers of compound rows (Table 1).

For an additional cross-validation test, we reduced the number of features to investigate their influence on the predictivity. *DB_c.50d*, *DB_c.25d*, and *DB_c.10d* contained reduced number of features, up to 50, 25, and 10%, respectively. *DB_c.8* contained only eight simplest features: AMW, average molecular weight; nSK, number of non-hydrogen atoms; snBO, number of non-hydrogen bonds; sRBN, number of rotatable bonds; nDB, number of double bonds; sMLOGP, Moriguchi octanol−water partition coefficient (logP); nHDon, number of hydrogen bond donors (N and O); and nHAcc, number of H-bond acceptors (N, O, F). *DB_c.1* was the unit feature vector with the sole feature equal to 1 for all compounds. Feature selection for *DB_c.50d*, *DB_c.25d*, and *DB_c.10d* data sets was performed in three random replicates to ensure the robustness of selection.

2.2.2. Viral Species Features. We used viral species (`species_id`) and genus (`genus_id`) information from the ViralChEMBL data set according to the ICTV 2014 taxonomy.[45,55] We exploited the genus assignment as the only (pseudo)feature of the species and represented it as a binary vector according to the following rule: if a species belonged to a genus, then the bit corresponding to the `genus_id` was set to 1, otherwise to 0. The resulting viral feature matrices *DB_v.main*, *DB_v.ext_points*, and *DB_v.ext_CS* contained 74 feature columns. It was not expected that such features should contain sufficient predictive information. The utilization of viral feature matrices was mandatory in the SGIMC method, while the models were designed with mostly compound feature-based prediction in mind.

2.3. Prediction Scenarios. Three main challenges can be addressed in the multitask prediction (Figure 2): prediction of new interactions for compounds and viruses, data for which were used for model building (Figure 2a), prediction of interaction

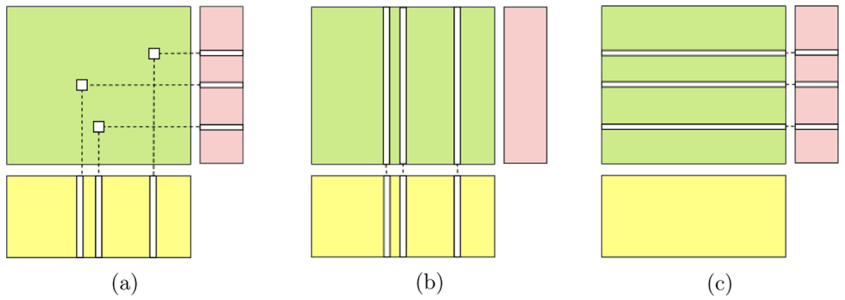


Figure 2. Addressed challenges: (a) prediction of point compound−virus interactions, (b) compoundwise CS prediction, and (c) specieswise CS prediction. Matrix of interactions, green; matrix of species features, pink; matrix of compound features, yellow; and unknown compound−virus interactions, white.

https://dx.doi.org/10.1021/acsomega.0c00857
ACS Omega 2020, 5, 15039−15051

profiles for compounds that were not used for model building (compoundwise CS prediction) (Figure 2b), and prediction of interactions for viral species that were not used for model building (specieswise CS prediction) (Figure 2c). We assessed the performance of the methods in all three scenarios, where possible.

CF algorithms process only the interaction matrix and thereby cannot perform CS predictions. Thus, Surprise CF algorithms were used to address only the prediction of new interactions between compounds and viruses utilized for model building (Figure 2a). Hyperparameters for the best models were selected by grid search (Supporting Information Table S1) and 10-fold cross-validation in the *model_selection* module. Models were based on the *DB_main* data set and external validation was performed on the *DB_ext_points* data set.

The SGIMC CBF algorithm was applied for solving all three challenges (Figure 2). Models were trained on the interaction matrix *DB_main* with side feature matrices *DB_c.main* and *DB_v.main*. External validation was performed using the interaction matrix *DB_ext_points* and the feature matrices *DB_c.ext_points* and *DB_v.ext_points* in the case of interaction prediction for known compounds and species and using matrices *DB_ext_CS*, *DB_c.ext_CS*, and *DB_v.ext_CS* in the case of compoundwise CS prediction. Model selection was performed based on a grid search of hyperparameters (Supporting Information Table S1) and cross-validation. We used *sklearn.model_selection* v.0.21.3 classes *KFold* for 10-fold cross-validation in the compoundwise CS scenario and *StratifiedKFold* for stratified 10-fold cross-validation of prediction of new interactions of known compounds and viruses due to substantial data set imbalance.

Models for solving specieswise CS problems were built without cross-validation using *DB_main*, *DB_c.main*, and *DB_v.main* data sets and hyperparameters of the best model for prediction of new interactions between known compounds and viral species. Model building and evaluation were performed by excluding the activity profiles of each species one by one from the model building and applying them as external test sets. Due to a small number of interaction values for the majority of species, the assessment of their prediction power could not be accurate.

2.3.1. Influence of the Number of Features. We carried out an additional test to investigate the influence of the number of features on the predictive power of the SGIMC algorithm. The model for the *DB_main* matrix was built. Feature matrices were represented by *DB_v.main* and *DB_c.main* subsets with different numbers of features, three generated by random samplings: *DB_c.50d*, *DB_c.25d*, and *DB_c.10d*, and two with fixed features: *DB_c.8* and *DB_c.1*. Assessment of prediction on reduced matrices was based on the stratified 10-fold cross-validation with hyperparameters from the best model for prediction of interaction for known compounds and viruses based on the original *DB_c.main* data set.

2.4. Evaluation and Metrics. Models were built based on the *DB_main* set and validated via cross-validation and external test sets (*DB_ext_CS* and *DB_ext_points*).[56,57] To avoid possible errors caused by the substantial imbalance of the data set with regard to activity classes, we used stratified 10-fold cross-validation keeping the constant proportion of active and inactive class assignments (90 and 10%, respectively) in the training and test sets.[58] We used grid search to optimize the hyperparameters of our algorithms. Varied hyperparameters and their values are shown in Supporting Information Table S1.

We used the receiver operating characteristic area under the curve (ROC AUC) score and two metrics based on it to assess prediction quality. The ROC AUC score was calculated using the *roc_auc_score* class of *sklearn.metrics* for all predicted values.[59] The mean and median of *n*-fold-averaged ROC AUC for a set of viral species[60] were also calculated

$$\text{mean/median ROCAUC} = \text{mean/median}\left\{\frac{1}{N}\sum_{n=1}^{N}\text{ROC AUC}_n(t) | t = 1, ..., T\right\} \quad (3)$$

where $\text{ROC AUC}_n(t)$ is the ROC AUC score calculated for viral species t for the test fold n, $N = 10$ in the case of cross-validation and $N = 1$ in the case of external validation. Standard deviation (SD) was calculated by *numpy.std* class (*numpy v.* 1.17.2).

We used the ROC AUC score to assess the prediction quality for all of the interaction values in the test set. The mean and median ROC AUC scores were used to demonstrate the difference in prediction quality for the separate viral species. For the comparison of models, we used median ROC AUC as the main measure not skewed by extremely large or small values, so it would better describe the real prediction quality. Also, SD was defined as a deviation for predictions for different viral species in the mean ROC AUC calculation and as a deviation in ROC AUC values for different cross-validation runs. The quality of specieswise CS prediction was assessed by the ROC AUC score for each species separately. ROC AUC, mean and median ROC AUC scores for the 10 best models for every algorithm are collected in Supporting Information Tables S2−S9. A code snippet for the metrics calculation is available as Supporting Information File SI2.

The robustness of the models was assessed by *y*-scrambling.[56,61] The median ROC AUC scores for normal models were compared with the median ROC AUC scores of the ones based on *y*-scrambled data sets. We used $1 - (n/m)$ as a measure of robustness,[62,63] where m and n are the numbers of "normal" and scrambled models with ROC AUC >0.6, respectively. The *y*-scrambling was performed for the 10 best models in each scenario according to the median ROC AUC score and was applied for both cross-validation and external validation (Supporting Information Tables S2−S9).

To assess the applicability of constructed models, we compared training and test data sets based on the similarity distance between their compounds. The distance was evaluated between *DB_c.main* and *DB_c.ext_points*, between *DB_c.main* and *DB_c.ext_CS*, and as an average of distances between training and test sets in every fold in the case of cross-validation (*DB_c.main*). The distance between each pair of compounds in the training and test data sets was computed based on their feature values. The cosine distance was calculated with the *spatial.distance.cdist* class of the *scipy* package. The similarity between the training and test sets was assessed based on the distribution of distances between every *i*th compound in a test set and all of the compounds in the training set (DIST_i), calculated according to the equation

$$\text{DIST}_i = \sum_{j=1}^{T}\frac{n_j}{N_i}\text{sim}_{ij} \quad (4)$$

where n_j is the number of known interactions of the *i*th compound of the test set and the *j*th compound of the training set with the same viral species, N_i is the overall number of known

https://dx.doi.org/10.1021/acsomega.0c00857
ACS Omega 2020, 5, 15039−15051

Table 2. Predictivity of Surprise Models

	cross-validation			external validation		
Surprise methods	ROC AUC ± SD	mean ROC AUC ± SD	median ROC AUC	ROC AUC	mean ROC AUC ± SD	median ROC AUC
knns.KNNBasic msd (virus-based)	0.808 ± 0.004	0.8 ± 0.3	0.86	0.603	0.7 ± 0.3	0.72
knns.KNNBasic msd (compound-based)	0.888 ± 0.004	0.8 ± 0.3	0.83	0.888	0.8 ± 0.3	0.83
knns.KNNBasic cosine (virus-based)	0.806 ± 0.005	0.8 ± 0.3	0.86	0.606	0.7 ± 0.3	0.75
knns.KNNBasic cosine (compound-based)	0.872 ± 0.004	0.7 ± 0.3	0.79	0.872	0.7 ± 0.3	0.75
co_clustering.CoClustering	0.863 ± 0.01	0.7 ± 0.3	0.81	0.702	0.7 ± 0.3	0.76
matrix_factorization.SVD	0.939 ± 0.003	0.8 ± 0.3	0.88	0.764	0.7 ± 0.3	0.78
matrix_factorization.NMF	0.939 ± 0.003	0.8 ± 0.3	0.88	0.709	0.7 ± 0.3	0.68

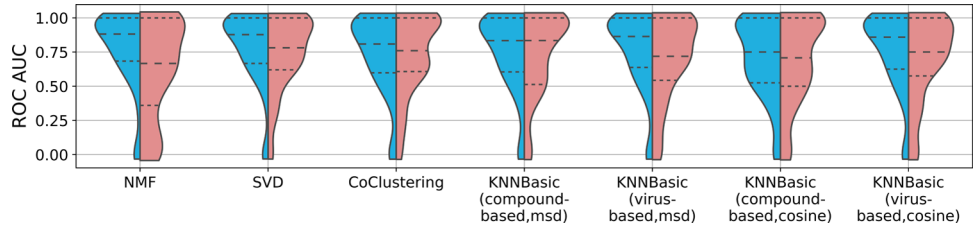


Figure 3. Violin plot of ROC AUC values for viral species in cross-validation (blue) and external validation (red). Dotted lines inside the violins represent the quartiles of the distribution.

interactions for the *i*th compound, and sim_{ij} is the cosine distance between the *i*th compound of the test set and the *j*th compound of the training set containing *T* compounds.

3. RESULTS AND DISCUSSION

The efficiency of antiviral activity class prediction with CF and CBF techniques was assessed and compared for small molecules from ViralChEMBL and ChEMBL 24. We represented compound−virus interactions as two classes, active and inactive, and encoded them in the interaction matrix as 1 and 0, respectively. In the case of a lack of experimental measurements, the corresponding value was kept empty. To understand the performance and robustness of the RS approaches, we investigated four scenarios:

- prediction of point interactions for known compounds and viruses,
- compoundwise CS prediction (prediction of interaction profiles for new compounds),
- specieswise CS prediction (prediction of interaction profiles for new viruses), and
- prediction of compound−virus interactions with a reduced number of compound features.

In all of the scenarios, the ROC AUC scores for the best models were much higher than the corresponding mean and median ROC AUC (Supporting Information Tables S2−S9). Thus, the ROC AUC scores could be used only to illustrate how precise the prediction was for all of the interaction values in the test set. It was calculated based on all predicted values and did not take into account the specifics of each viral species. At the same time, the mean/median ROC AUC was the average/median of ROC AUC values separately calculated for each viral species. A moderate mean/median ROC AUC value and a high SD of the mean ROC AUC indicated satisfactory prediction for the whole data set along with a substantial difference in prediction power for different viral species: the activity class prediction based on the same model for some viral species was perfect, while for the others it was unsatisfactory.

We performed 10-fold cross-validation and optimized hyperparameters of models by grid search. To prove the lack of impact of data set imbalance on the prediction results, the *y*-scrambling test was performed for the 10 best models under each scenario in both cross-validation and external validation settings (Supporting Information Tables S2−S9). Upon *y*-randomization, the quality of models decreased, providing compelling evidence of the relevance of our prediction model.

We examined the opportunity of using the Balanced Accuracy and the Precision-Recall AUC as model quality metrics.[59] Their use led to overestimation of prediction results:[59] for the best models, their values were equal to 0.98−0.99. Thus, we did not use them for the model assessment. We also did not assess the accuracy of our models because it is easy to get high accuracy even for a poor model for an imbalanced data set.[64] We did not use specificity, sensitivity/recall, precision, median balanced accuracy, and similar metrics that are suitable for imbalanced data because these metrics require an active/inactive threshold, which is not applicable for a data set based on several targets.

We also evaluated the similarity of data sets by comparing the distance from each compound in the test set to all of the compounds in the training sets. The results are presented in Supporting Information Figure S1 and Table S10. External test sets were found to consist of compounds that are more distant from the training set compounds compared with the compounds in the training and test sets during cross-validation.

3.1. Prediction of Point Interactions. *3.1.1. Collaborative Filtering.* We explored three collaborative filtering techniques: *k*-nearest neighbors, coclustering, and matrix factorization. The performance of the best models is given in Table 2 and is illustrated in Figure 3.

It should be noted that the cross-validation prediction results in Surprise suffer from the CS problem. Due to a high data set sparsity, more than 60% of the compounds possess only one interaction (it is about 40% of compound−species interactions in the *DB_main*). Predicted values for these compound−virus pairs during cross-validation will be equal to the mean of all interactions from their viral species profile.

https://dx.doi.org/10.1021/acsomega.0c00857
ACS Omega 2020, 5, 15039−15051

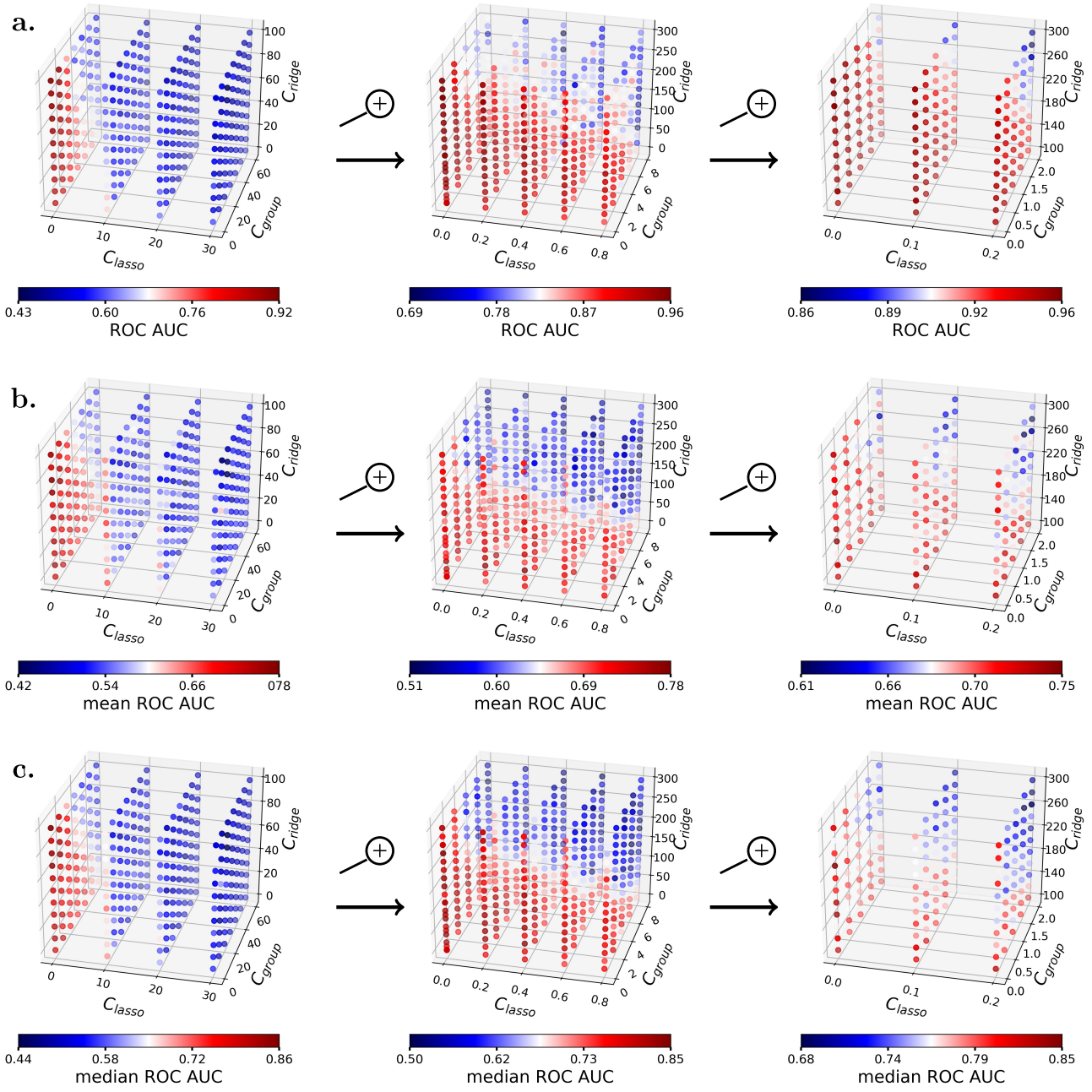


Figure 4. Guided grid search of C_{lasso}, C_{ridge}, and C_{group} coefficients for interaction prediction for known compounds and viral species based on (a) ROC AUC, (b) mean ROC AUC, and (c) median ROC AUC. Rank = 10, number of iterations = 70.

KNNBasic methods are directly derived from the *k*-nearest-neighbors approach and follow the basic paradigm of chemo-informatics: similar compounds possess similar properties. In our case, this statement can also be extended as follows: similar compounds interact with similar viruses and similar viruses are inhibited by similar compounds. The similarity is calculated for the interaction profiles of compounds or viruses. The performance of models varies depending on the similarity metric as well as the direction of similarity calculation: virus- or compound-based similarity. Compound-based models demonstrate better predictive power (Table 2). However, the similarity calculation is both the key factor and the bottleneck of this algorithm. Upon an increase of the number of interaction profiles N, the predictive power of the model increased, probably due to the increase of the information capacity of the similarity matrix, but at the same time, space and time complexity is $O(N^2)$. It makes the applicability of similarity-based CF methods limited for large data sets. For example, the *DB_main* data set required at most 1.5 GB RAM and a dozen seconds for the calculation of an msd or cosine similarity matrix for 158 viral species. The same data set required at least 1700 GB RAM and 2 h or 2500 GB and 6 h for an *msd* or cosine calculation of 250K compounds, respectively.

Methods based on coclustering and matrix factorization do not rely on profile similarity; therefore, they do not need large RAM resources (no more than 1.5 GB of RAM) and take several minutes under the same computational conditions as those for *KNNBasic* methods. In the coclustering, rows and columns of an interaction matrix are simultaneously grouped to compare the profiles and complete the missing values. The best coclustering

https://dx.doi.org/10.1021/acsomega.0c00857
ACS Omega 2020, 5, 15039−15051

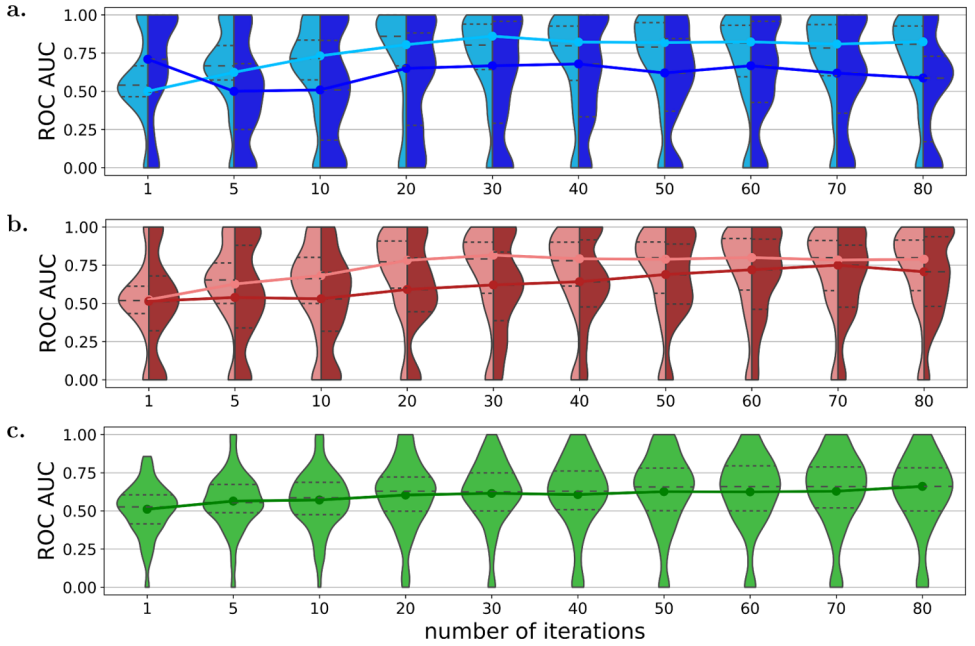


Figure 5. Violin plots of ROC AUC values for viral species: (a) prediction of point compound−virus interactions, (b) compoundwise CS prediction, and (c) specieswise CS prediction. The prediction was assessed in cross-validation (light blue and coral) and external validation (dark blue, red, and green). Lines depict the dependence of median ROC AUC scores on the number of iterations. Dotted lines inside the violins represent the quartiles of the distribution. Rank = 10, C_{lasso} = 0.0, C_{group} = 0.0, and C_{ridge} = 120.0.

model shows a cross-validation median ROC AUC of 0.81, which is between the cosine virus-based kNN (0.79) and compound-based kNN (0.86). Both compound- and virus-based *msd* also perform better than coclustering in the cross-validation (median ROC AUCs of 0.83 and 0.86). For the test set, the median ROC AUC is almost the same for coclustering and kNN methods (around 0.75), except for compound-based *msd* (0.83). Thus, coclustering may be used in place of kNN if the computational resources are limited.

The matrix factorization approach solves the problem of matrix completion by finding latent features that determine the internal relationship in data (in our case, between compounds and viruses). Models based on this approach showed the best performance in the 10-fold cross-validation protocol: median ROC AUC = 0.88 in both cases. On the test set, the NMF models demonstrated the worst performance (median ROC AUC = 0.68), while the prediction power of the SVD models (median ROC AUC = 0.78) is second only to the *msd* compound-based kNN model (median ROC AUC = 0.83).

3.1.2. CBF Prediction of Point Interactions with SGIMC. The problem of matrix completion is considered as an optimization procedure using the features of compounds and viruses. The SGIMC algorithm shares the idea of the IMC approach of matrix completion by combining feature vectors, associated with row and column entities of the interaction matrix, with a low-rank matrix. Three matrices are required to train an SGIMC model: a partially filled interaction matrix and full feature matrices for compounds and viruses. Our compound feature matrix *DB_c.main* was filled with Dragon descriptors, whereas the virus feature matrix *DB_v.main* included only genus assignments. By design, SGIMC has an option for feature selection, which is implemented through a sparsity-inducing penalty and its regularization coefficient C_{group}. Also, coefficients C_{ridge} and C_{lasso}, representing the squared Frobenius norm and the matrix L_1-norm penalties, respectively, are involved in regularization. These regularization coefficients were varied along with the rank of the internal low-rank matrix **W** and the number of training iterations to choose the best SGIMC model.

For our best model (median ROC AUC = 0.84), we have the following hyperparameters: rank = 10, number of iterations = 70, C_{lasso} = 0.0, C_{ridge} = 120.0, and C_{group} = 0.0. An increase of C_{lasso} and C_{group} leads to a notable decrease of performance, while an increase of C_{ridge} leads to its slight increase, followed by a slow descent (Figure 4) after the optimal C_{ridge} value of about 120.

3.2. Cold-Start Prediction with CBF. The cold-start problem is a possible lack of performance of a recommender system applied to a new compound or virus, for which there is no experimental data. In particular, the problem is critical for collaborative filtering methods, based on the interaction matrix only. To tackle this issue, CBF approaches (e.g., SGIMC), based on available side-channel information (features of compounds/viruses), may be used to obtain reliable predictions in the cold-start mode.

We established the hyperparameters for the best SGIMC model for the compoundwise CS prediction based on a cross-validation grid search. The compoundwise CS performance appeared not to differ substantially from prediction for known compounds and viruses in the quality of interaction prediction (Figure 5). For example, the models with one of the highest predictive powers (median ROC AUCs of 0.82 and 0.86, respectively) were built on the same hyperparameter set: rank = 10, number of iterations = 70, C_{lasso} = 0.0, C_{ridge} = 120.0, and C_{group} = 0.0. Test set efficiency of the model based on these hyperparameter values was assessed by median ROC AUC, which was equal to 0.71 and 0.69 for compoundwise CS prediction and point interaction prediction for known compounds and viruses, respectively (Figure 5a,b). A substantial decrease in predictive quality for external test sets is a result of differences between their compounds and compounds in the training set.

SGIMC models for specieswise CS prediction based on the interaction matrix *DB_main* and feature matrices *DB_c.main*

https://dx.doi.org/10.1021/acsomega.0c00857
ACS Omega 2020, 5, 15039−15051

and *DB_v.main* appeared to be inferior to the models for compoundwise CS prediction (Figure 5b,c) (Supporting Information Tables S3 and S4). In specieswise CS prediction, the median value of ROC AUC for all viral species is 0.65 at 70 iterations, while it is 0.75 in the case of compoundwise CS prediction on the external test set (rank = 10, C_{lasso} = 0.0, C_{ridge} = 120.0, C_{group} = 0.0). Moreover, the distribution of median ROC AUC values for new compounds (Figure 5b) is more shifted toward the 1 than the distribution of ROC AUC values for new species (Figure 5b). For example, the top quartile is more than 0.9 for compoundwise CS prediction and more than 0.8 for specieswise CS prediction, while the bottom quartile is close to 0.5 in both cases.

The decrease of the predictive power in the specieswise CS was apparently caused by the insufficient virus features, represented by the genus assignment only. To prove this hypothesis, we modeled the situation with absolutely uninformative features by replacing *DB_c.main* and *DB_v.main* matrices with unit vectors (Figure 6). It is clear that the models

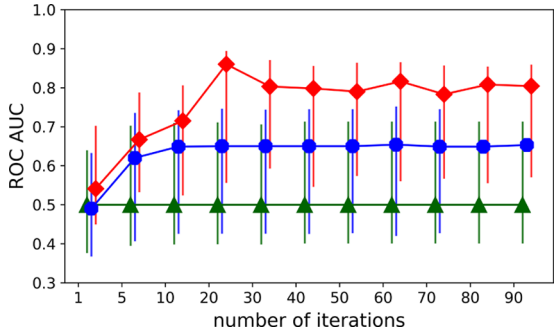


Figure 6. Dependence of the median ROC AUC score for point interaction prediction on number of iterations through cross-validation with original feature matrices (red), unit vector for species (blue), and unit vector for compounds (green) (rank = 10, C_{lasso} = 0.0, C_{group} = 0.0, and C_{ridge} = 120.). Error bars represent the SD.

based on the original feature matrices possess the best predictive quality (Figure 6, red). In the case of unit vector for species (Figure 6, blue), models were based on sufficient information from compound features and were still able to predict interaction values, though with lower prediction quality. In the case of unit vector for compounds (Figure 6, green), interaction prediction was not meaningful, indicating that the original species features are not sufficient. We hope that the proper introduction of virus features will improve the results for all scenarios.

3.3. Influence of the Number of Features. The compound feature information in the *DB_c.main* matrix is redundant, so we supposed that the features could be randomly removed without a significant deterioration in prediction quality. To assess this hypothesis, we carried out a series of experiments for known compounds and viruses by replacement of the *DB_c.main* matrix with truncated feature matrices *DB_c.50d*, *DB_c.25d*, *DB_c.10d*, *DB_c.8*, and *DB_c.1*.

This experiment showed that compound feature information in *DB_c.main* is excessive for the SGIMC models (Figure 7). Reducing the feature number up to 50, 25, and 10% continuously, but slightly, reduced the prediction quality of the models. For the models built on the *DB_c.main*, *DB_c.50d*, *DB_c.25d*, and *DB_c.10d* sets, the median ROC AUC scores were 0.81, 0.78, 0.74, and 0.72, respectively (rank = 10, number of iterations = 60, C_{lasso} = 0.0, C_{ridge} = 120.0, C_{group} = 0.0). With the same hyperparameters, the models based on the eight simplest features demonstrated a critical decrease of the prediction quality, with the median ROC AUC = 0.67. The models based on unit vectors were not predictive at all.

SGIMC allows one to select features using the C_{group} penalty coefficient to filter out the noninformative ones. The selection of the most significant features is performed by the increase of the C_{group} coefficient: with its increase, the number of selected features is decreased. In the SGIMC authors' benchmarks, the method was able to select about 6000 features from the set of 355 709.[52] However, in our experiment, the increase of C_{group} led to a decrease of the mean/median ROC AUC values and, thereby, the predictive quality of a model. The extent of the decrease depended on the other hyperparameters (Figure 8).

It is clear from the comparison of Figures 7 and 8 that the models based on the same number of selected compound features possess different predictive powers depending on whether the features were selected randomly (Figure 7) or using the C_{group} coefficient (Figure 8). For example, in cross-validation for point interaction prediction with 50% of the compound features, ROC AUC scores were 0.79 for the models with random selection and 0.68 for the models with C_{group} selection (number of iterations = 70, rank = 10, C_{lasso} = 0.0, C_{ridge} = 120.0). It was a result of an application of the C_{group} coefficient for both compound and species feature selections, i.e., using C_{group} led to zeroing of both compound and species features simultaneously. The strategy of a simultaneous feature selection is smart in the case of a huge amount of noisy features, which is not the case in our task, characterized by the insufficiency of

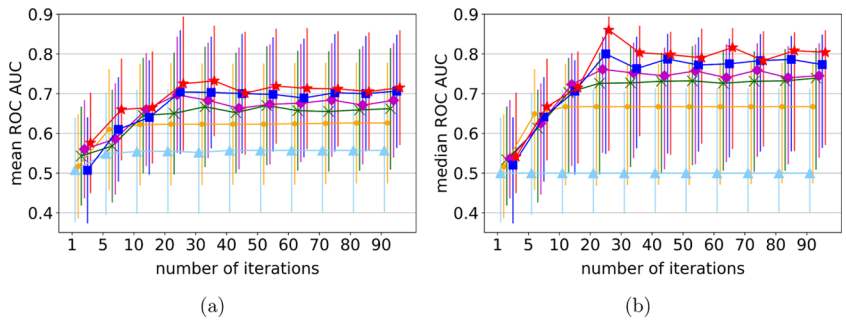


Figure 7. Dependence of mean ROC AUC (a) and median ROC AUC (b) for models with a different number of compound features on the number of iterations. Rank = 10, C_{lasso} = 0.0, C_{group} = 0.0, and C_{ridge} = 120.0. Compound feature matrices: *DB_c.main* (red ★), *DB_c.50d* (blue ■), *DB_c.25d* (magenta ◆), *DB_c.10d* (green ×), *DB_c.8* (orange •), and *DB_c.1* (light blue ▲). Error bars represent the SD.

https://dx.doi.org/10.1021/acsomega.0c00857
ACS Omega 2020, 5, 15039−15051

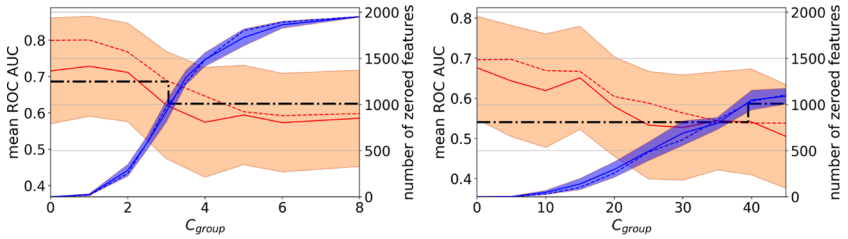


Figure 8. Influence of the C_{group} regularization coefficient in cross-validation for point interaction prediction on the mean/median ROC AUC at 70 (a) and 10 (b) iterations. Continuous and dashed red lines indicate the mean and median ROC AUC, and continuous and dashed blue lines indicate the mean and median number of zeroed features. Shaded areas represent the corresponding standard deviations. The black dash-dotted line shows median ROC AUC with 50% of compound features. C_{lasso} = 0.0, C_{ridge} = 120.0, and rank = 10.

species features. It led to a critical loss of feature information and deterioration of model quality. Separate determination of regularization coefficients for both compound and species feature matrices should be a solution to this problem.

4. CONCLUSIONS

Multitask prediction algorithms have been gaining ground rapidly with the appearance of databases storing multitarget data. The recommender system (RS) as an approach of multitask prediction may be a powerful tool for compound−target interaction prediction. These methods allow one to predict the activity class for all combinations of compounds and targets in a data set and select the best of them for further experimental investigations. However, the current experience in this domain is limited and far from complete.

Our experiments demonstrated that RS algorithms based on collaborative and content-based filtering to a sparse matrix of antiviral activity data can achieve sustainable performance for the antiviral activity class prediction. We revealed that both approaches showed a very high predictive ability in cross-validation and external validation as measured by ROC AUC and mean/median ROC AUC.

Collaborative filtering (CF) methods demonstrate high performance but they possess several crucial limitations. The models based on the calculation of compound profile similarity demonstrate the best predictive ability among the investigated CF methods but the application of these methods is challenging due to the requirement of a huge amount of RAM for the similarity calculation and storage. Improvement of the algorithm by reducing the required RAM during model building would allow the wider use of these methods for data sets with thousands of compounds. The matrix factorization methods lead to models with moderate predictive ability. Their preference over other CF methods is determined by the simplicity of their application. The main disadvantage of all CF methods is a limited applicability domain: we can make a prediction only for compounds or viruses whose interaction profiles were used during model creation.

The application of content-based filtering (CBF) algorithms is preferable because of the possibility of using feature information for compounds and viruses. Using compounds' features, CBF makes the cold-start prediction possible. The main disadvantage of the approach is the requirement of generation and processing of additional feature information, which can be a challenging task in the case of viruses, and may require a lot of computational resources. The SGIMC method allows one to reduce the number of used features to several thousand, which was not possible in our case with only 2016 features. Without feature selection, the SGIMC algorithm was virtually reduced to the IMC. Using this algorithm, we demonstrated that the prediction of antiviral activity for both new and known compounds against known viruses can be performed with rather high accuracy, while prediction of the antiviral activity of known compounds against new viruses was less accurate due to the insufficient characterization of the viruses in our data set. We believe that the development of appropriate virus features could solve the problem; however, it may be a tricky issue by itself.

This research revealed promising applicability and effectiveness of the RS approaches in drug discovery. We hope that further progress in this field can be achieved with hybrid RS approaches that can make the best of CF and CBF models.

APPENDIX: TERMINOLOGICAL ISSUES IN THE RS FIELD

RS approaches have evolved rapidly in the field of e-commerce and have found extensive application in other fields. Although their application was common, there was a tendency to study these approaches in different communities under different names. Therefore, despite the fact that the term "recommender systems" has begun to appear in drug discovery research only in the last few years,[65] methods based on them began to be exploited around a decade ago.[36,66−68] The clearest example is the proteochemometric approach[8,9] and "read across" methods.[11] Also, there is a terminological confusion even within the same community. For example, the matrix of multitarget multicompound interaction data with columns corresponding to the compounds and rows corresponding to the targets may be called "interaction matrix",[69,70] "interaction space",[9] or "target matrix".[71]

This terminological inconsistency led to the lack of RS terms in the publications related to drug discovery. Analysis of the publishing activity by related queries ("recommender systems", "recommendation systems", "collaborative filtering", "content-based filtering", and "drug") in the title, abstract, and keywords in Scopus and Dimensions[72,73] showed that interest in the application of RS for drug discovery and design has appeared since 2012. Nevertheless, algorithms, on which RS are based, have been applied to this field for decades and have generated greater interest.

To avoid confusion in terminology, in this article, multitarget data are represented and defined as interaction matrix **M**. Commonly,[13] it may be represented by triplets (x_i, y_j, m_{ij}) of instances $x \in \mathbf{X}$, targets $y \in \mathbf{Y}$, and scores of their relationship $m \in \mathbf{M}$. In drug discovery, triplets are represented by compounds, their biological targets, and their *interactions*, defined as the

https://dx.doi.org/10.1021/acsomega.0c00857
ACS Omega 2020, 5, 15039−15051

activity of the compounds against these targets. The vector of interactions for each target or compound forms an interaction profile. Interactions may be represented by nominal, ordinal, or real values. In the drug discovery context, they may correspond to activity values (e.g., IC_{50}, K_i) or activity classes (e.g., boolean: 0, inactive; 1, active; and multiclass: 3, high activity; 2, moderate activity; 1, weak activity; 0, inactive). Scores of compound−target interaction generated by an RS are referred to as recommended values.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.0c00857.

File SI1: pdf file with a description of data set preparation; table of models' hyperparameters; and the best hyperparameters' and prediction assessment (PDF)

File SI2: python code snippet for the metric calculation; File SI3: gzipped tarball file with the data sets inside (*DB_main.csv*—data set of compound−virus interactions, *DB_c_main.csv*—data set of compound features, *DB_v_main.csv*—data set of virus features, *DB_ext.csv*—test data set with compound−virus interactions, *DB_c_ext.csv*—test data set with compound features, and *DB_ext_comp.csv* —data set with compound labels for point and CS test prediction); file is located on Zenodo (doi: 10.5281/zenodo.3831446)

■ AUTHOR INFORMATION

Corresponding Author

Ekaterina A. Sosnina − *Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, Moscow 143026, Russia; Institute of Physiologically Active Compounds, RAS, Chernogolovka 142432, Russia;* orcid.org/0000-0002-6764-755X; Phone: +79260256249; Email: ekaterina.sosnina@skoltech.ru

Authors

Sergey Sosnin − *Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, Moscow 143026, Russia; Syntelly LLC, Skolkovo Innovation Center, Moscow 121205, Russia;* orcid.org/0000-0002-3042-7369

Anastasia A. Nikitina − *Department of Chemistry, Lomonosov Moscow State University, Moscow 119991, Russia; FSBSI "Chumakov FSC R&D IBP RAS", Moscow 108819, Russia*

Ivan Nazarov − *Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, Moscow 143026, Russia*

Dmitry I. Osolodkin − *FSBSI "Chumakov FSC R&D IBP RAS", Moscow 108819, Russia; Institute of Translational Medicine and Biotechnology, Sechenov First Moscow State Medical University, Moscow 119991, Russia;* orcid.org/0000-0002-0462-2945

Maxim V. Fedorov − *Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, Moscow 143026, Russia; Syntelly LLC, Skolkovo Innovation Center, Moscow 121205, Russia; Physics John Anderson Building, University of Strathclyde, Glasgow G4 0NG, U.K.*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c00857

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the usage of the Skoltech CDISE HPC clusters Arkuda and Zhores for obtaining the results presented in this manuscript. The authors are thankful to Maxim Panov and Evgeny Frolov from the Center for Computational and Data-Intensive Science and Engineering, Skoltech, for fruitful discussions. The reported study was funded by the Russian Foundation of Basic Research (according to the research project no. 19-33-90290, MVF and EAS—computational experiments and results assessment) and the State research funding for FSBSI "Chumakov FSC R&D IBP RAS" (topic no. 0837-2019-0002, AAN and DIO—database curation and data assessment).

■ NOMENCLATURE

Acronyms

RS	recommender system
CF	collaborative filtering
CBF	content-based filtering
CS	cold-start
SGIMC	sparse-group inductive matrix completion
SD	standard deviation

■ REFERENCES

(1) Caruana, R. Multitask Learning. *Mach. Learn.* **1997**, *28*, 41−75.

(2) Lipinski, C. F.; Maltarollo, V. G.; Oliveira, P. R.; da Silva, A. B. F.; Honorio, K. M. Advances and Perspectives in Applying Deep Learning for Drug Design and Discovery. *Front. Robot. AI* **2019**, *6*, 108.

(3) Norinder, U.; Svensson, F. Multitask Modeling with Confidence Using Matrix Factorization and Conformal Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 1598−1604.

(4) Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **2019**, *5*, No. eaav6490.

(5) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma. *J. Chem. Inf. Model.* **2017**, *57*, 2068−2076.

(6) Sosnin, S.; Vashurina, M.; Withnall, M.; Karpov, P.; Fedorov, M.; Tetko, I. A Survey of Multi-task Learning Methods in Chemoinformatics. *Mol. Inf.* **2018**, 615−621.

(7) Sosnin, S.; Karlov, D.; Tetko, I. V.; Fedorov, M. V. Comparative Study of Multitask Toxicity Modeling on a Broad Chemical Space. *J. Chem. Inf. Model.* **2019**, 1062−1072.

(8) van Westen, G. J. P.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric Modeling as a Tool to Design Selective Compounds and for Extrapolating to Novel Targets. *MedChemComm* **2011**, *2*, 16−30.

(9) Cortés-Ciriano, I.; Ain, Q. U.; Subramanian, V.; Lenselink, E. B.; Méndez-Lucio, O.; IJzerman, A. P.; Wohlfahrt, G.; Prusis, P.; Malliavin, T. E.; van Westen, G. J. P.; Bender, A. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *MedChemComm* **2015**, *6*, 24−50.

(10) Schaduangrat, N.; Anuwongcharoen, N.; Phanus-umporn, C.; Sriwanichpoom, N.; Wikberg, J. E.; Nantasenamat, C. *Silico Drug Design*; Elsevier, 2019; pp 281−302.

(11) Alves, V. M.; Golbraikh, A.; Capuzzi, S. J.; Liu, K.; Lam, W. I.; Korn, D. R.; Pozefsky, D.; Andrade, C. H.; Muratov, E. N.; Tropsha, A. Multi-Descriptor Read Across (MuDRA): A Simple and Transparent Approach for Developing Accurate Quantitative Structure-Activity Relationship Models. *J. Chem. Inf. Model.* **2018**, *58*, 1214−1223.

(12) *Recommender Systems Handbook*; Ricci, F.; Rokach, L.; Shapira, B., Eds.; Springer US: Boston, MA, 2015.

(13) Waegeman, W.; Dembczyński, K.; Hüllermeier, E. Multi-target prediction: a unifying view on problems and methods. *Data Min. Knowl. Discovery* **2019**, *33*, 293−324.
(14) Bennett, J.; Elkan, C.; Liu, B.; Smyth, P.; Tikk, D. KDD Cup and Workshop 2007. *SIGKDD Explor. Newsl.* **2007**, *9*, 51−52.
(15) Amatriain, X.; Basilico, J. *Recommender Systems Handbook*; Ricci, F.; Rokach, L.; Shapira, B., Eds.; Springer US: Boston, MA, 2015; pp 385−419.
(16) Thorat, P. B.; Goudar, R. M.; Barve, S. Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System. *Int. J. Comput. Appl.* **2015**, *110*, 31−36.
(17) Aggarwal, C. C. *Recommender Systems: The Textbook*, 1st ed.; Springer Publishing Company, Inc., 2016.
(18) Sanghavi, B.; Rathod, R.; Mistry, D. M. Recommender Systems-Comparison of Content-based Filtering and Collaborative Filtering. *Int. J. Curr. Eng. Technol.* **2014**, *4*, 3131−3133.
(19) Aggarwal, P.; Tomar, V.; Kathuria, A. Comparing Content Based and Collaborative Filtering in Recommender Systems. *Int. J. New Technol. Res.* **2017**, *3*, 3.
(20) Ariff, N. M.; Bakar, M. A. A.; Rahim, N. F. In *Comparison Between Content-based and Collaborative Filtering Recommendation System for Movie Suggestions*, AIP Conference Proceedings; AIP Publishing LLC: Kuala Lumpur, Malaysia, 2018; p 020057.
(21) Su, X.; Khoshgoftaar, T. M. A Survey of Collaborative Filtering Techniques. *Adv. Artif. Intell.* **2009**, *2009*, 1−19.
(22) Nilashi, M.; Bagherifard, K.; Ibrahim, O.; Alizadeh, H.; Nojeem, L. A.; Roozegar, N. Collaborative Filtering Recommender Systems. Res. *J. Appl. Sci., Eng. Technol.* **2013**, *5*, 4168−4182.
(23) Sharma, M.; Mann, S. A Survey of Recommender Systems: Approaches and Limitations. *Int. J. Innov. Sci. Eng. Technol.* **2013**, 1−9.
(24) Cacheda, F.; Carneiro, V.; Fernández, D.; Formoso, V. Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Trans. Web* **2011**, *5*, 1−33.
(25) Kumar Bokde, D.; Girase, S.; Mukhopadhyay, D. Matrix Factorization Model in Collaborative Filtering Algorithms: A Survey. *Procedia Comput. Sci.* **2015**, *49*, 136−146.
(26) Pazzani, M. J.; Billsus, D. *The Adaptive Web*; Brusilovsky, P.; Kobsa, A.; Nejdl, W., Eds.; Springer: Berlin, Heidelberg, 2007; Vol. *4321*, pp 325−341.
(27) Lops, P.; de Gemmis, M.; Semeraro, G. *Recommender Systems Handbook*; Ricci, F.; Rokach, L.; Shapira, B.; Kantor, P. B., Eds.; Springer US: Boston, MA, 2011; pp 73−105.
(28) Zhang, W.; Zou, H.; Luo, L.; Liu, Q.; Wu, W.; Xiao, W. Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing* **2016**, *173*, 979−987.
(29) Fan, J.; Yang, J.; Jiang, Z. Prediction of Central Nervous System Side Effects Through Drug Permeability to Blood-Brain Barrier and Recommendation Algorithm. *J. Comput. Biol.* **2018**, *25*, 1−9.
(30) Wang, H.; Gu, Q.; Wei, J.; Cao, Z.; Liu, Q. Mining drug-disease relationships as a complement to medical genetics-based drug repositioning: Where a recommendation system meets genome-wide association studies. *Clin. Pharmacol. Ther.* **2015**, *97*, 451−454.
(31) Hao, W.; Hai-ping, W.; Xin-dong, W.; Qi, L. Mining Drug-Disease Relationships: a Recommendation System. *Chin. Pharmacol. Bull.* **2015**, *31*, 1770−1774.
(32) Yang, J.; Li, Z.; Fan, X.; Cheng, Y. Drug-Disease Association and Drug-Repositioning Predictions in Complex Diseases Using Causal Inference-Probabilistic Matrix Factorization. *J. Chem. Inf. Model.* **2014**, *54*, 2562−2569.
(33) Galeano, D.; Paccanaro, A. A Recommender System Approach for Predicting Drug Side Effects. In *International Joint Conference on Neural Networks (IJCNN)*; IEEE, 2018; pp 1−7.
(34) Qiu, H.; Mao, K.-T.; Shi, J.-Y.; Huang, H.; Chen, Z.; Dong, K.; Yiu, S.-M. Predicting and Understanding Comprehensive Drug-Drug Interactions via Semi-nonnegative Matrix Factorization. *BMC Syst. Biol.* **2018**, 101−110.
(35) Shi, J.-Y.; Huang, H.; Li, J.-X.; Lei, P.; Zhang, Y.-N.; Yiu, S.-M. Predicting Comprehensive Drug-Drug Interactions for New Drugs via Triple Matrix Factorization. *Bioinf. Biomed. Eng.* **2017**, *2018*, 108−117.
(36) Yamada, M.; Lian, W.; Goyal, A.; Chen, J.; Wimalawarne, K.; Khan, S. A.; Kaski, S.; Mamitsuka, H.; Chang, Y. In *Convex Factorization Machine for Toxicogenomics Prediction*, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '17, pp 1215−1224.
(37) Bhat, S.; Aishwarya, K. In *Item-Based Hybrid Recommender System For Newly Marketed Pharmaceutical Drugs*, 2013 International Conference on Advances in Computing, Mysore, 2013; pp 2107−2111.
(38) Huang, Z.; Lu, X.; Duan, H.; Zhao, C. Collaboration-based Medical Knowledge Recommendation. *Artif. Intell. Med.* **2012**, *55*, 13−24.
(39) Ma, J.; Zhang, R.; Yuan, Y.; Zhao, Z. Using Hybrid Similarity-Based Collaborative Filtering Method for Compound Activity Prediction. In *International Conference on Intelligent Computing*; Springer: Cham, 2018; pp 51−72.
(40) Simm, J.; Arany, A.; Zakeri, P.; Haber, T.; Wegner, J. K.; Chupakhin, V.; Ceulemans, H.; Moreau, Y. In *Macau: Scalable Bayesian Factorization with High-Dimensional Side Information Using MCMC*, 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing, 2017; pp 2107−2111.
(41) de León, A.; Chen, B.; Gillet, V. J. Effect of Missing Data on Multitask Prediction Methods. *J. Cheminf.* **2018**, *10*, No. 26.
(42) Hasan, S.; Duncan, G. T.; Neill, D. B.; Padman, R. In *Towards a Collaborative Filtering Approach to Medication Reconciliation*, AMIA Annual Symposium Proceedings, 2008; pp 288−292.
(43) Hasan, S.; Duncan, G. T.; Neill, D. B.; Padman, R. Automatic Detection of Omissions in Medication Lists. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 449−458.
(44) Huang, Z.; Lu, X.; Duan, H.; Zhao, C. Collaboration-based Medical Knowledge Recommendation. *Artif. Intell. Med.* **2012**, *55*, 13−24.
(45) Nikitina, A. A.; Orlov, A. A.; Kozlovskaya, L. I.; Palyulin, V. A.; Osolodkin, D. I. Enhanced Taxonomy Annotation of Antiviral Activity Data from ChEMBL. *Database* **2019**, *2019*, No. bay139.
(46) Seley-Radtke, K. L.; Yates, M. K. The Evolution of Nucleoside Analogue Antivirals: A Review for Chemists and Non-chemists. Part 1: Early Structural Modifications to the Nucleoside Scaffold. *Antiviral Res.* **2018**, *154*, 66−86.
(47) Yates, M. K.; Seley-Radtke, K. L. The Evolution of Antiviral Nucleoside Analogues: A Review for Chemists and Non-chemists. Part II: Complex Modifications to the Nucleoside Scaffold. *Antiviral Res.* **2019**, *162*, 5−21.
(48) Li, G.; De Clercq, E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). *Nat. Rev. Drug Discov.* **2020**, *19*, 149−150.
(49) Grčar, M.; Mladenič, D.; Fortuna, B.; Grobelnik, M. *Advances in Web Mining and Web Usage Analysis*; Hutchison, D.; Kanade, T.; Kittler, J.; Kleinberg, J. M.; Mattern, F.; Mitchell, J. C.; Naor, M.; Nierstrasz, O.; Pandu Rangan, C.; Steffen, B., Eds.; Springer: Berlin, Heidelberg, 2006; Vol. *4198*, pp 58−76.
(50) Guo, M. *User Modeling, Adaptation, and Personalization*; Hutchison, D.; Kanade, T.; Kittler, J.; Kleinberg, J. M.; Mattern, F.; Mitchell, J. C.; Naor, M.; Nierstrasz, O.; Pandu Rangan, C.; Steffen, B., Eds.; Springer: Berlin, Heidelberg, 2012; Vol. *7379*, pp 361−364.
(51) Hug, N. Surprise, a Python Library for Recommender Systems. http://surpriselib.com (accessed December 1, 2018).
(52) Nazarov, I.; Shirokikh, B.; Burkina, M.; Fedonin, G.; Panov, M. Sparse Group Inductive Matrix Completion, 2018. arXiv preprint arXiv:1804.10653. https://arxiv.org/abs/1804.10653.
(53) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **2015**, *43*, W612−W620.
(54) Kode srl, Dragon (Software for Molecular Descriptor Calculation), version 7.0.8., 2017. https://chm.kode-solutions.net.
(55) ICTV Master Species List, v.1. https://talk.ictvonline.org/files/master-species-lists/m/msl/5945 (accessed July 1, 2018).

https://dx.doi.org/10.1021/acsomega.0c00857
ACS Omega 2020, 5, 15039−15051

(56) Muhammad, U.; Uzairu, A.; Ebuka Arthur, D. Review on: quantitative structure activity relationship (QSAR) modeling. *J. Anal. Pharm.* **2018**, *7*, 240−242.
(57) Siontis, G. C.; Tzoulaki, I.; Castaldi, P. J.; Ioannidis, J. P. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J. Clin. Epidemiol.* **2015**, *68*, 25−34.
(58) Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, 2018. arXiv preprint arXiv:1811.12808. https://arxiv.org/abs/1811.12808.
(59) Brown, J. B. Classifiers and their Metrics Quantified. *Mol. Inform.* **2018**, *37*, 1−11.
(60) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively Multitask Networks for Drug Discovery, 2015. arXiv preprint arXiv:1502.02072. https://arxiv.org/abs/1502.02072.
(61) Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345−2357.
(62) Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Feng, J.; Zheng, W.; Tropsha, A. QSAR Modeling of Datasets with Enantioselective Compounds using Chirality Sensitive Molecular Descriptors. *SAR QSAR Environ. Res.* **2005**, *16*, 93−102.
(63) de Cerqueira Lima, P.; Golbraikh, A.; Oloff, S.; Xiao, Y.; Tropsha, A. Combinatorial QSAR Modeling of P-Glycoprotein Substrates. *J. Chem. Inf. Model.* **2006**, *46*, 1245−1254.
(64) Tharwat, A. Classification assessment methods. *Appl. Comput. Inform.* **2018**, *12*, No. e0177678.
(65) Ozsoy, M. G.; Özyer, T.; Polat, F.; Alhajj, R. Realizing Drug Repositioning by Adapting a Recommendation System to Handle the Process. *BMC Bioinform.* **2018**, *19*, 263−266.
(66) Yang, J.; Li, Z.; Fan, X.; Cheng, Y. Drug-Disease Association and Drug-Repositioning Predictions in Complex Diseases Using Causal Inference-Probabilistic Matrix Factorization. *J. Chem. Inf. Model.* **2014**, *54*, 2562−2569.
(67) Ding, H.; Takigawa, I.; Mamitsuka, H.; Zhu, S. Similarity-based Machine Learning Methods for Predicting Drug-Target Interactions: a Brief Review. *Brief. Bioinform.* **2014**, *15*, 734−747.
(68) Martin, E. J.; Polyakov, V. R.; Zhu, X.-W.; Mukherjee, P.; Tian, L.; Liu, X. All-Assay-Max2 pQSAR: Activity Predictions as Accurate as 4-concentration IC50s for 8558 Novartis Assays. *J. Chem. Inf. Model.* **2019**, *59*, 4450−4459.
(69) Koohi, A. In *Prediction of Drug-Target Interactions Using Popular Collaborative Filtering Methods*, 2013 IEEE International Workshop on Genomic Signal Processing and Statistics, 2013; pp 58−61.
(70) Peska, L.; Buza, K.; Koller, J. Drug-Target Interaction Prediction: A Bayesian Ranking Approach. *Comput. Methods Programs Biomed* **2017**, *152*, 15−21.
(71) Ezzat, A.; Zhao, P.; Wu, M.; Li, X.-L.; Kwoh, C.-K. Drug-Target Interaction Prediction with Graph Regularized Matrix Factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *14*, 646−656.
(72) Science, D. Data Sourced from Dimensions, an Inter-linked Research Information System Provided by Digital Science. https://www.dimensions.ai (accessed October 1, 2018).
(73) Hook, D. W.; Porter, S. J.; Herzog, C. Dimensions: Building Context for Search and Evaluation. *Front. Res. Metrics Anal.* **2018**, *3*, 23.

https://dx.doi.org/10.1021/acsomega.0c00857
ACS Omega 2020, 5, 15039−15051